The attention illusion: overcoming the localization-prediction discrepancy in biomass ViT-based regression

Farida Nchare[1,2,3], Norbert Tsopze[2,3], Corina Iovan[1]

[1] Institut de Recherche pour le Développement, UMR ENTROPIE, Noumea, BP A5 - 98848 cedex, New Caledonia
[2] Department of Computer Science, University of Yaounde I, Yaounde, P.O. Box 812, Cameroon
[3] Sorbonne University, IRD, UMMISCO, F-93143, Bondy, France

Abstract

Biomass estimation from images is important for ecological monitoring but relies on morphometric features that are not always visually salient. We observe that while Vision Transformers (ViTs) produce attention maps suggesting precise spatial localization, this visual focus does not guarantee accurate weight prediction. Using a frozen DINO-based ViT for bivalve biomass estimation, we show that morphometric information is primarily encoded in intermediate layers rather than final semantic representations. Applying learnable multi-head attention pooling over these layers significantly improves regression performance, achieving an $R^2 = 0.950$ and reducing RMSE by 39.6% over standard CLS aggregation. These results reveal an attention illusion in ViTs and demonstrate the critical role of intermediate features for high-precision morphometric regression.

Keywords

Vision Transformers, Attention analysis, Feature aggregation, Biomass estimation, Explainability.

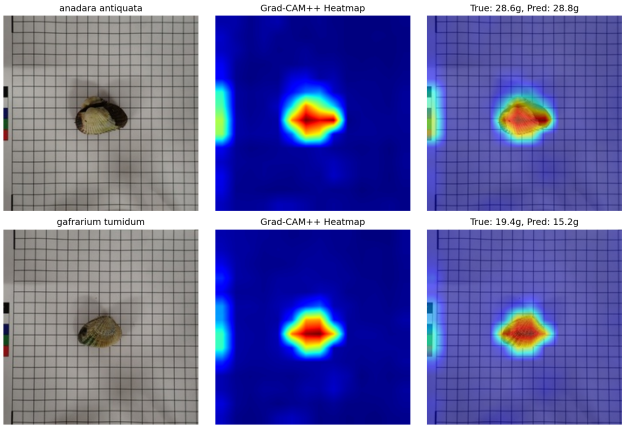

Figure 1: DINO self-attention maps for two bivalve specimens from different species. The model predicts 28.8 g for a 28.6 g specimen (top) but 15.2 g for a 19.4 g specimen (bottom). Despite similar localization patterns, prediction accuracy differs significantly, illustrating the *attention illusion*.

1 Introduction

Estimating biomass directly from images enables non-invasive monitoring of marine species populations across ecosystems. Unlike classification tasks that focus on categorical identity, biomass estimation relies on *morphometric information*, meaning structural properties such as size, curvature, and volumetric features that correlate with organism mass. Extracting these continuous geometric signals from images is challenging, particularly when specimens exhibit strong cross-species variability. Individuals with similar visual appearance may differ in density, making accurate mass estimation dependent on subtle morphological features.

Vision Transformers (ViTs) trained with self-supervised objectives such as DINO [3] produce attention maps that sharply delineate object boundaries without task-specific supervision. This spatial precision suggests that such models should capture the geometric structure required for morphometric estimation. However, standard pipelines extract the global representation from the final-layer `CLS` token, a design inherited from classification models [5]. Classification and regression impose fundamentally different representational requirements. DINO encourages semantic invariance by training the `CLS` token to produce similar embeddings for augmented views of the same image, preserving object identity while discarding variations in scale or geometry. In contrast, biomass regression requires these variations to remain encoded in the representation. As transformer layers progressively transform low-level features into higher-level semantics [9], fine-grained morphometric information may therefore be compressed or lost. Consequently, as illustrated in Figure 1, two specimens can exhibit nearly identical attention patterns while producing very different prediction accuracy. We refer to this mismatch between spatial localization and predictive performance as the *attention illusion*.

This work investigates where morphometric information is encoded within the ViT hierarchy. Through layer-wise regression analysis, we show that predictive performance peaks at intermediate transformer layers rather than at the final semantic layer. Motivated by this observation,

we propose Adaptive Multi-head Attention Pooling (A-MAP), a feature aggregation strategy that operates on patch tokens from these intermediate layers instead of relying on the global CLS representation. A-MAP builds upon MAP [7] and uses learnable attention to aggregate spatial features, enabling different heads to capture complementary geometric structures. To further enrich the representation, the method aggregates tokens from adjacent intermediate layers, providing a multi-level description of specimen geometry.

We evaluate the proposed approach on a dataset of more than 2000 images of bivalves from two species (*Anadara antiquata* and *Gafrarium tumidum*) with weights ranging from 0.2 g to 66.6 g. Using a frozen DINO ViT-S/16 backbone, the model is trained without species labels so that predictions must rely on morphometric cues rather than species identity. Our analysis reveals that morphometric information peaks at intermediate layers, and aggregating these representations with A-MAP achieves $R^2 = 0.950$, reducing RMSE by 39.6% compared to standard CLS aggregation.

Our contributions are as follows:

- We identify the *attention illusion* in ViT-based regression, showing that accurate spatial attention does not guarantee accurate morphometric prediction.
- We conduct a systematic layer-wise analysis demonstrating that morphometric information peaks at intermediate transformer layers rather than at the final semantic representation.
- We propose A-MAP, an attention-based pooling strategy that aggregates intermediate-layer patch tokens and improves by $+10.0\%$ regression performance over standard CLS aggregation.

2 Related Work

ViT representation hierarchy. Vision Transformers (ViTs) exhibit a progressive evolution of representations across their depth, transitioning from localized spatial patterns to abstract semantic descriptors [9, 5]. Early layers retain fine-grained textures and edge information, whereas deeper layers increasingly encode class-level abstractions [3]. Self-supervised frameworks such as DINO demonstrate that attention maps can highlight coherent object regions without explicit supervision [3]. Despite this, most downstream applications still rely exclusively on the final-layer [CLS] token as a global image representation [5]. Recent studies suggest that spatial relationships and object structure are distributed across the transformer feature space rather than concentrated in a single token [8]. However, the representational hierarchy of ViTs is largely shaped by classification objectives. As a result, the class-discriminative information emphasized in the deepest layers may not preserve the continuous spatial structure required for precise morphometric regression.

Feature aggregation beyond CLS. To address the limitations of the [CLS] token, several aggregation strategies have been proposed to better exploit patch tokens. Multi-head Attention Pooling (MAP) introduces learnable queries that attend to spatial tokens, enabling a more flexible representation than fixed global pooling [7]. Other approaches focus on multi-scale feature fusion. For instance, HAFA [4] and SDPT [2] combine representations from multiple transformer layers to capture both low-level textures and high-level semantic features. In regression tasks, architectures such as FASNet [11] have successfully employed shallow and deep features jointly for crowd density estimation. Nevertheless, these multi-scale strategies typically assume that aggregating more features leads to better representations, often performing uniform fusion across layers. Such aggregation may dilute task-relevant information, as signals useful for morphometric estimation could be overlooked by the increasing semantic abstraction of deeper layers. This suggests the need for more selective strategies that explicitly target layers where morphometric features are most informative.

Morphometric regression from images. Image-based estimation of biomass or weight has become an important tool for ecological monitoring [13]. Prior work has developed effective pipelines for fish biomass estimation [12], segmentation-based morphometric analysis [6], and joint detection–estimation frameworks for species-specific prediction [10]. However, these methods typically extract features from the final backbone layer. Whether applied to fish or marine invertebrates [1], they implicitly assume that the deepest representation is optimal for regression. Yet this assumption has rarely been examined for morphometric estimation. This work addresses this gap by investigating the intermediate feature space of ViTs and proposing a regression pipeline designed to extract morphometric information from these layers.

3 Method

We investigate where morphometric information resides in frozen Vision Transformer (ViT) representations and how to exploit it for regression. The method consists of three stages: (1) problem formulation, (2) layer selection via layer-wise regression evaluation, and (3) feature aggregation using Adaptive Multi-head Attention Pooling (A-MAP).

3.1 Problem Setup

Given an RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the goal is to predict a scalar weight $\hat{y} \in \mathbb{R}^{+}$ (in grams). We use a frozen DINO ViT-S/16 backbone [3]. For each image, the transformer produces patch token embeddings at each layer l:

$$\mathbf{X}_l \in \mathbb{R}^{N\times D},$$

where $N = 196$ is the number of patch tokens, $D = 384$ the embedding dimension, and $l \in \{0, \ldots, L-1\}$ with $L = 12$ layers. The backbone remains frozen throughout training. Only the pooling and regression layers are optimized.

Standard practice extracts the CLS token from the final layer as a global representation. However, the *attention*

illusion suggests that late-layer representations may discard the morphometric information needed to predict mass. Our objective is therefore to learn a pooling function

$$f : \mathbf{X}_l \to \mathbf{z}$$

that extracts a fixed-size representation $\mathbf{z} \in \mathbb{R}^D$, which is then mapped to the scalar prediction $\hat{y}$ by a regression head. The central question is twofold: at which depth l does the backbone encode the geometric features most useful for mass regression, and how should we aggregate patch tokens to exploit them?

3.2 Layer Selection Protocol

For tasks that depend on geometric features, intermediate representations may preserve more task-relevant information than the final semantic layer. To test whether this representational mismatch reflects a depth-dependent phenomenon, we evaluate regression performance independently at each transformer layer. For each layer $l \in \{0, \ldots, L-1\}$, we extract the patch token representation $\mathbf{X}_l$ and train an independent regression head g_l. Each model uses the same pooling and regression architecture to ensure fair comparison across layers. Validation performance is then used to identify the optimal layer

$$l^* = \arg\max_l R^2(g_l(\mathbf{X}_l)).$$

This procedure is architecture-agnostic with respect to the regression head, as any pooling mechanism can serve as g_l. The specific head used in our analysis is described in Section 4.

3.3 Multi-head Attention Pooling

Multi-head Attention Pooling (MAP) [7] replaces the fixed `CLS` token with a learnable aggregation over patch tokens. A learnable query attends to the token sequence using standard multi-head scaled dot-product attention, producing a pooled representation in $\mathbb{R}^D$. This mechanism allows the model to emphasize informative spatial regions rather than relying on the fixed `CLS` representation.

While MAP provides effective features aggregation, it operates on tokens from a single layer. The layer selection protocol may reveal a zone of high performance around the peak, suggesting that combining adjacent layers could capture geometric features at complementary levels of abstraction. A-MAP extends MAP to exploit this structure.

3.4 Adaptive Multi-head Attention Pooling (A-MAP)

To exploit the representational structure revealed by the layer selection protocol, A-MAP attends jointly over patch tokens from multiple intermediate layers centered on the morphometric peak. We denote H the number of attention heads and $d_k = D/H$ the per-head dimension. A-MAP operates through two mechanisms: multi-layer concatenation and a peak-based attention bias.

Multi-layer concatenation. Let l_1, l_2, l_3 denote three layers selected by the protocol. This window captures the morphometric peak and its immediate neighborhood, balancing representation richness against computational cost. We concatenate rather than fuse their representations, as concatenation preserves each layer's token identity and lets the attention mechanism learn which depth is most informative per spatial location:

$$\mathbf{X}_{\text{concat}} = [\mathbf{X}_{l_1}; \mathbf{X}_{l_2}; \mathbf{X}_{l_3}] \in \mathbb{R}^{3N \times D} \tag{1}$$

where $[;]$ denotes row-wise concatenation. This yields $3N$ tokens of dimension D, forming a single sequence that spans three levels of the representational hierarchy.

The remaining question is how to bias attention toward the most informative layer without hard-coding the selection.

Attention with peak-based bias. A learnable query $\mathbf{q}_h \in \mathbb{R}^{1 \times d_k}$ for each head h attends over keys $\mathbf{k}_j \in \mathbb{R}^{1 \times d_k}$ derived from $\mathbf{X}_{\text{concat}}$. We augment the standard scaled dot-product score with a learnable per-layer bias:

$$e_j = \frac{\mathbf{q}_h \, \mathbf{k}_j^\top}{\sqrt{d_k}} + b_{l(j)} \tag{2}$$

where $l(j)$ denotes the source layer of token j in the concatenated sequence. The term $b_{l(j)} \in \mathbb{R}$ is a learnable scalar bias assigned to all tokens from layer $l(j)$. The bias for the peak layer l^* is initialized to $+2.0$, while the others are initialized to 0.0. This initialization encourages early attention toward the morphometrically optimal layer while remaining fully learnable during training.

The per-head attended output is computed as:

$$\mathbf{z}_h = \text{softmax}(\mathbf{e})\, \mathbf{V}_h \in \mathbb{R}^{1 \times d_k} \tag{3}$$

where $\mathbf{V}_h \in \mathbb{R}^{3N \times d_k}$ contains the value projections of $\mathbf{X}_{\text{concat}}$ for head h. Each head produces a weighted summary of the most morphometrically relevant tokens across all three layers. The H head outputs are concatenated and projected to $\mathbb{R}^D$ via a linear layer.

The complete pipeline is summarized as:

$$\hat{y} = g\big(\text{A-MAP}\big([\mathbf{X}_{l_1}; \mathbf{X}_{l_2}; \mathbf{X}_{l_3}]\big)\big)$$

3.5 Regression Head

The pooled representation $\mathbf{z} \in \mathbb{R}^D$ is passed to a lightweight MLP that outputs the scalar prediction $\hat{y}$. The pooling module and regression head are trained jointly, while the backbone remains frozen. Implementation details are provided in Section 4.

4 Experiments

We evaluate whether morphometric information required for biomass estimation is preserved in the standard Vision Transformer pipeline. Our experiments address three questions: (1) where geometric information is encoded within the ViT hierarchy, (2) whether the commonly used `CLS` representation retains these features, and (3) whether spatial aggregation over patch tokens can recover them. Using a frozen DINO ViT-S/16 backbone, we perform a layer-wise regression analysis and compare multiple

feature aggregation strategies. The analysis reveals a clear morphometric peak at intermediate layers where regression accuracy is substantially higher than when using the final-layer CLS token. Motivated by this observation, we evaluate Adaptive Multi-head Attention Pooling (A-MAP), which aggregates intermediate patch representations using learnable attention.

4.1 Experimental Setup

Dataset. We evaluate on the *Bivalves* dataset, comprising 2406 RGB images of marine bivalves from three species: *Anadara antiquata*, *Gafrarium tumidum*, and *Gafrarium pectinatum*. Individual weights range from 0.2 g to 66.6 g. We adopt a cross-species evaluation with model trained and tested on the data regardless of the species identity.

Evaluation Metrics. We report four regression metrics: coefficient of determination (R^2), root mean squared error (RMSE, in grams), mean absolute error (MAE, in grams), and mean absolute percentage error (MAPE). All metrics are reported with 95% bootstrap confidence intervals ($N = 1000$ resamples).

Implementation Details. We use a frozen DINO ViT-S/16 backbone with 12 transformer layers, producing 196 patch tokens of dimension 384 per image. The regression head consists of three fully connected layers ($384 \to 256 \to 128 \to 1$) with dropout. We train with AdamW (lr $= 0.001$, wd $= 0.01$) and a ReduceLROnPlateau scheduler (factor $= 0.8$, patience $= 10$). The loss function is MSE. Batch size is 16, maximum epochs 50, with early stopping (patience $= 15$). All experiments use seed 42.

4.2 Locating Morphometric Information in ViT Representations

Before proposing a new aggregation strategy, we investigate where morphometric information is encoded within the ViT hierarchy. We extract patch tokens from each transformer layer and evaluate regression performance using MAP pooling. Figure 2 shows the resulting R^2 curve across layers.

The curve exhibits a clear inverted U-shape with three regimes. Layer 0 corresponds to raw patch embeddings and yields limited predictive performance ($R^2 = 0.593$). Performance rises rapidly in early layers as contour and structural features emerge.

Regression accuracy peaks in intermediate layers (3–6), with the maximum at layer 5 and nearly identical performance at layer 6. We refer to this region as the *morphometric peak*, where geometric descriptors such as shell curvature and growth patterns are best preserved.

Beyond this point performance gradually declines, reflecting the increasing semantic abstraction induced by the DINO objective [3]. The geometric detail required for mass estimation therefore resides primarily in intermediate representations.

Two conclusions follow. First, the standard CLS pipeline relies on representations where morphometric information has already been partially compressed. Second, intermediate patch tokens retain geometric features that can be exploited by spatial aggregation methods.

4.3 Main Results

Table 1 compares different aggregation strategies using the same frozen DINO ViT-S/16 backbone. Results show that performance depends on both the aggregation mechanism and the representation depth.

The CLS baseline extracts a single global vector from the final layer, discarding spatial structure. Adding attention weighting (CLS-Att.) or concatenating CLS tokens from multiple layers (CLS Multiscale) yields moderate improvements but remains limited.

Replacing fixed aggregation with learnable attention produces a larger gain. MAP pooling introduces a learnable query that attends over patch tokens, improving R^2 from 0.864 to 0.926 on the same layer 11 features.

A further improvement arises from selecting the appropriate representation depth. Layer 5 MAP outperforms layer 11 MAP by +2.1 points despite identical pooling. Selecting the appropriate representation depth therefore has a larger impact on morphometric regression performance than the choice of pooling mechanism.

A-MAP combines both insights by applying multi-head attention over patch tokens concatenated from layers $[4, 5, 6]$, centered on the morphometric peak. This achieves the best performance across all metrics, reducing RMSE by 39.6% compared to the CLS baseline.

4.4 Visual Analysis of Predictions

Figure 3 shows predicted versus ground-truth biomass for the CLS baseline and A-MAP. A-MAP predictions cluster closely around the identity line across the full weight range, while the CLS baseline shows systematic underestimation for heavier specimens. Preserving spatial morphometric information from intermediate layers therefore reduces bias, particularly for high-mass individuals where geometric features are most discriminative.

This observation relates to the *attention illusion* introduced in Section 1. Although DINO self-attention precisely localizes the shell contour (Figure 1), CLS-based regression performs substantially worse. Accurate spatial attention does not guarantee that the aggregated representation preserves morphometric information. The CLS token compresses patch features into a single vector optimized for the pretraining objective, causing geometric features useful for regression to be partially lost.

Figure 4 visualizes the spatial attention patterns of individual heads within the A-MAP pooling module. Each head attends to different shell regions, confirming that multi-head attention captures complementary morphometric features.

4.5 Ablation Studies

Effect of Aggregation Strategy. Table 2 compares aggregation strategies on identical layer 11 features. Mean pooling performs worst, indicating that uniform averaging discards important spatial information. Attention-weighted pooling improves performance but remains limited by

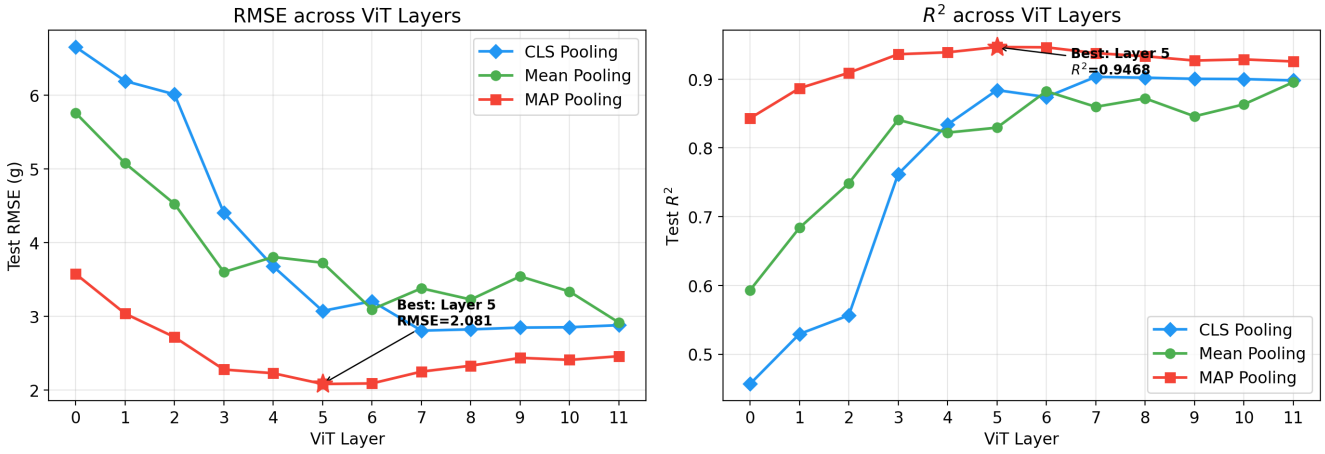

Figure 2: R^2 curve across all DINO ViT-S layers using MAP pooling. The morphometric peak occurs at layer 5 ($R^2 = 0.947$), indicated by ★. Early layers (0–2) encode low-level features; intermediate layers (3–6) encode geometric structure; late layers (7–11) show semantic compression.

Table 1: Comparison of pooling methods on the Bivalves dataset (test split). Best results in **bold**. All metrics reported with 95% bootstrap confidence intervals ($N = 1000$).

Method	Aggregation	Layers	R^2	RMSE (g)	MAE (g)	MAPE (%)
CLS (baseline)	CLS	11	0.864 ± 0.014	3.33 ± 0.10	2.51 ± 0.07	15.0 ± 0.7
CLS-Att.	CLS-Att.	11	0.910 ± 0.009	2.70 ± 0.08	2.03 ± 0.06	11.8 ± 0.6
CLS Multiscale	CLS	[5,6,7,8]	0.910 ± 0.009	2.71 ± 0.12	2.05 ± 0.08	12.0 ± 0.5
MAP	MAP	11	0.926 ± 0.008	2.46 ± 0.08	1.81 ± 0.06	10.4 ± 0.5
Multiscale MAP	MAP	[5,6,7,8]	0.943 ± 0.006	2.15 ± 0.07	1.63 ± 0.06	10.0 ± 0.5
Layer 5 MAP	MAP	5	0.947 ± 0.005	2.08 ± 0.08	1.57 ± 0.06	9.7 ± 0.5
A-MAP (ours)	**A-MAP**	**[4,5,6]**	$\mathbf{0.950 \pm 0.005}$	$\mathbf{2.01 \pm 0.09}$	$\mathbf{1.50 \pm 0.06}$	$\mathbf{8.6 \pm 0.4}$

fixed weights. The CLS-Attention hybrid further gains by incorporating spatial features alongside the `CLS` token. MAP pooling yields the best result, confirming that learnable task-specific attention provides the most effective aggregation.

Table 2: Ablation: aggregation strategy (Layer 11).

Method	R^2	RMSE (g)
Mean	0.826 ± 0.015	3.76 ± 0.17
Att-Weighted	0.879 ± 0.013	3.14 ± 0.17
CLS	0.864 ± 0.014	3.33 ± 0.10
CLS-Att.	0.910 ± 0.009	2.70 ± 0.08
MAP	0.926 ± 0.008	2.46 ± 0.08

Effect of Representation Depth. Table 3 evaluates MAP pooling at different depths. Performance peaks at layer 5, with nearly identical performance at layer 6. Intermediate layers outperform late layers by more than two R^2 points, confirming that representation depth plays a critical role in morphometric regression. This result aligns with the layer-wise analysis in Figure 2, which shows that geometric information is concentrated in intermediate transformer representations [3, 9].

Effect of Multi-layer Fusion. Table 4 compares multi-layer strategies. CLS Multiscale on $[5, 6, 7, 8]$ concatenates `CLS` tokens from multiple layers but achieves only $R^2 = 0.910$, comparable to the single-layer CLS-Att. baseline. Replacing `CLS` concatenation with MAP pooling on the same layers produces a clear gain (+4%), confirming that learnable attention is essential for exploiting multi-layer features. A-MAP narrows the layer window to $[4, 5, 6]$, centered on the morphometric peak, and achieves the best result.

Table 3: Ablation: representation depth (MAP, single layer).

Layer	R^2	RMSE (g)
11	0.926 ± 0.008	2.46 ± 0.08
8	0.933 ± 0.006	2.34 ± 0.08
6	0.946 ± 0.005	2.10 ± 0.08
5	$\mathbf{0.947 \pm 0.005}$	$\mathbf{2.08 \pm 0.08}$

Table 4: Ablation: multi-layer fusion strategy.

Method	Layers	R^2	RMSE (g)
CLS Multiscale	[5,6,7,8]	0.910 ± 0.009	2.71 ± 0.12
Multiscale MAP	[5,6,7,8]	0.943 ± 0.006	2.15 ± 0.07
A-MAP (ours)	**[4,5,6]**	$\mathbf{0.950 \pm 0.005}$	$\mathbf{2.01 \pm 0.09}$

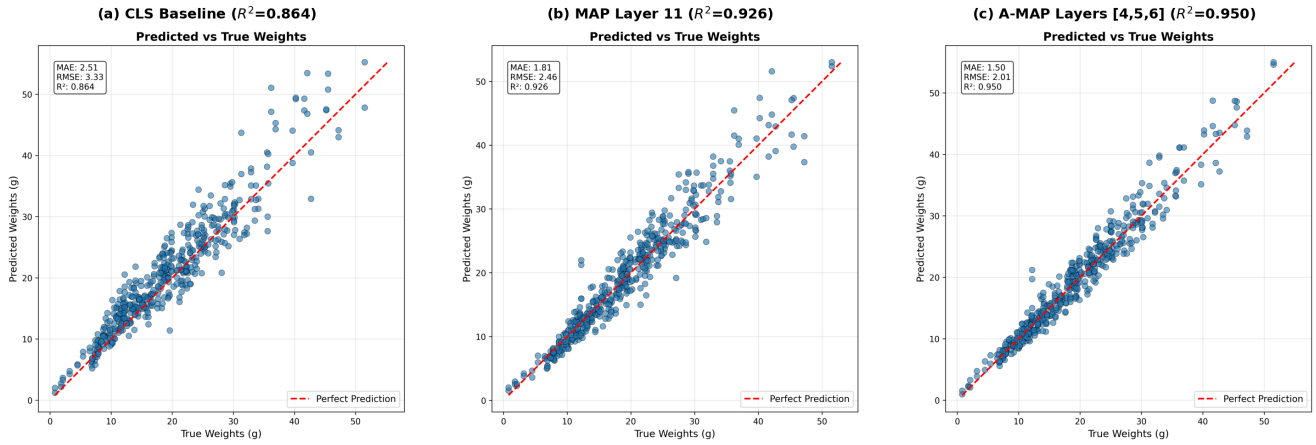

Figure 3: Predicted vs. actual weight scatter plots for the CLS baseline (Layer 11, top row) and A-MAP (bottom row). Columns show all species, *Anadara antiquata*, and *Gafrarium* spp., respectively. A-MAP predictions align closely with the identity line across the full weight range.

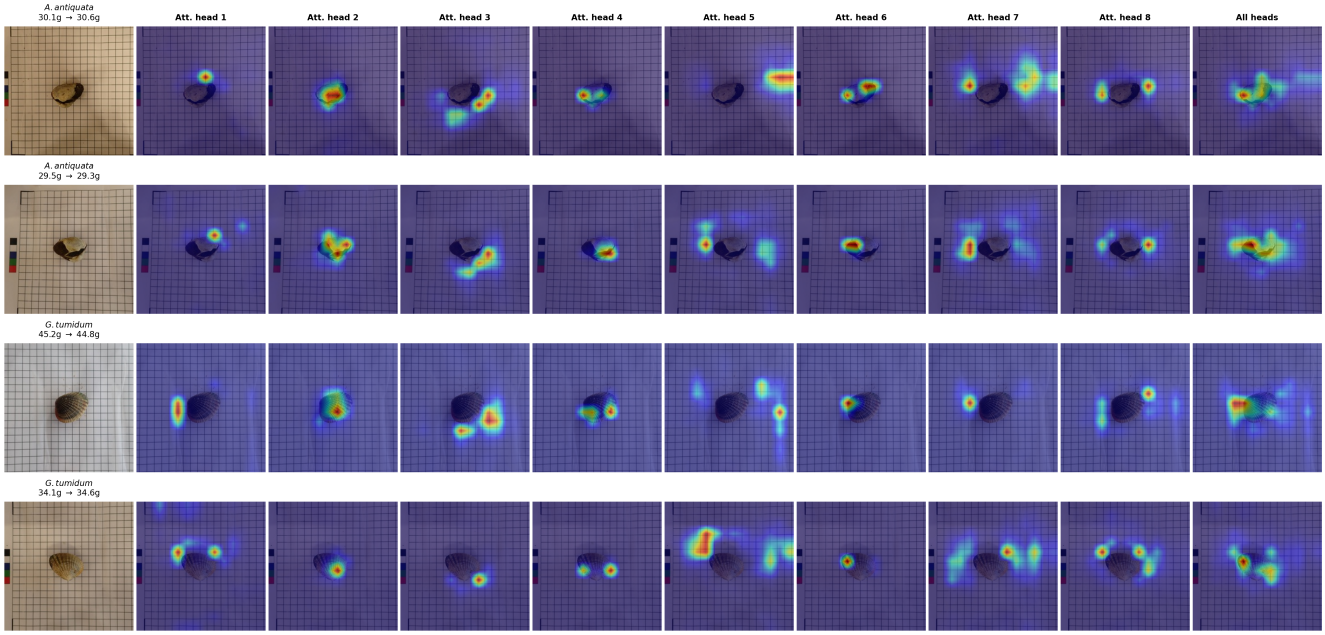

Figure 4: Spatial attention maps from the 8 attention heads of the A-MAP pooling module for two *Anadara antiquata* (top) and two *Gafrarium tumidum* (bottom) specimens. Each column corresponds to one learnable attention head attending over patch tokens from layers $[4, 5, 6]$. The rightmost column shows the aggregated attention across all heads. Different heads specialize in distinct morphometric structures.

4.6 Limitations

- We evaluate on a single morphological group (marine bivalves); generalization to other species or object categories remains untested.
- Our analysis uses a single backbone (DINO ViT-S/16); the morphometric peak layer may shift across architectures or pretraining objectives.
- All experiments use a frozen encoder; fine-tuning the backbone could redistribute morphometric information across layers.
- The dataset size is moderate (2406 images), which may limit the stability of deeper regression architectures.

5 Conclusion

In this work, we identified an *attention illusion* in DINO ViT: self-attention maps precisely localize objects, yet the CLS representation discards the geometric detail needed for weight regression. A systematic layer-wise analysis reveals that morphometric information peaks at intermediate depth (layer 5), not at the final semantic layer. A-MAP, a

learnable attention pooling over patch tokens from layers $[4, 5, 6]$ with peak-centric bias, achieves $R^2 = 0.950$. Compared to the CLS baseline, this reduces RMSE by 39.6% ($3.33\,\mathrm{g}$ to $2.01\,\mathrm{g}$) and MAPE by 42.7% (15.0% to 8.6%). Combining intermediate geometric features with late semantic context via gated fusion and validating across other physical property regression tasks remain promising directions for future work.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014241 made by GENCI.

Data availability

The dataset used during the current study is available in the public repository Zenodo 10.5281/zenodo.17936025.

References

[1] Johanna Ärje, Claus Melvad, Mads Jeppesen, Sigurd Madsen, Jenni Raitoharju, Maria Rasmussen, Alexandros Iosifidis, Ville Tirronen, Kristian Meissner, Moncef Gabbouj, and Toke Høye. Automatic image-based identification and biomass estimation of invertebrates. *Methods in Ecology and Evolution*, 11(8):922–931, 2020.

[2] Hu Cao, Guang Chen, Hengshuang Zhao, Dongsheng Jiang, Xiaopeng Zhang, and Qi Tian. SDPT: Semantic-aware dimension-pooling transformer for image segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):15934–15946, 2024.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[4] Yongjie Chen, Hongmin Liu, Haoran Yin, and Bin Fan. Building vision transformers with hierarchy aware feature aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5908–5917, 2023.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[6] Arthur F.A. Fernandes, Eduardo M. Turra, Érika R. de Alvarenga, Tiago L. Passafaro, Fernando B. Lopes, Gabriel F.O. Alves, Vikas Singh, and Guilherme J.M. Rosa. Deep learning image segmentation for extraction of fish body measurements and prediction of body weight and carcass traits in Nile tilapia. *Computers and Electronics in Agriculture*, 170:105274, 2020.

[7] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant set input. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753. PMLR, 2019.

[8] Saebom Leem and Hyunseok Seo. Attention guided CAM: Visual explanations of vision transformer guided by self-attention. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[9] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pages 12116–12128, 2021.

[10] Maria Sokolova, Manuel Cordova, Henk Nap, Aloysius van Helmond, Michiel Mans, Arjan Vroegop, Angelo Mencarelli, and Gert Kootstra. An integrated end-to-end deep neural network for automated detection of discarded fish species and their weight estimation. *ICES Journal of Marine Science*, 80(7):1911–1922, 2023.

[11] Kehao Wang, Yuhui Wang, Ruiqi Ren, Han Zou, and Zhichao Shao. Transformer-based feature aggregation and stitching network for crowd counting. *IEEE Access*, 11:124833–124844, 2023.

[12] Tianye Zhang, Yuqiao Yang, Yueyue Liu, Chenglei Liu, Ran Zhao, Daoliang Li, and Chen Shi. Fully automatic system for fish biomass estimation based on deep neural network. *Ecological Informatics*, 79:102399, 2024.

[13] Yuliang Zhao, Qijun Xiao, Jinhao Li, Kaixuan Tian, Le Yang, Peng Shan, Xiaoyong Lv, Lianjiang Li, and Zhikun Zhan. Review on image-based animals weight weighing. *Computers and Electronics in Agriculture*, 215:108456, 2023.